一种新型的负载均衡-交叉点缓冲交换结构

徐 宁[1,2],余少华[1,2],汪学舜[2]
(1.华中科技大学计算机学院,湖北武汉 430074;2.武汉邮电科学研究院新一代光纤通信技术和网络国家重点实验室,湖北武汉 430074)

摘 要: 针对混合输入-交叉点队列(CICQ)交换结构受限于"流控通信延时"、"需要 2 倍内部加速仿真输出队列(OQ)交换"以及单纯交叉点缓冲(CQ)存在"非均衡流量模式下吞吐量性能不足"等问题,本文提出一种新型的"负载均衡交叉点缓冲交换结构".采用固定模式时隙轮转匹配进行负载均衡处理,将到达输入端口的非均衡流量转化为近似均衡流量并且平均分配到同一输出端口对应的交叉缓冲中,从而可以利用较小的交叉点缓冲来模拟输出队列调度,简化调度过程并且提高吞吐量.理论分析证明了这种新结构的稳定性以及模拟输出队列交换的能力.同时仿真表明,采用该交换结构可以在不需要内部加速的条件下获得相当于输出队列交换的性能,并且有效地解决了交叉点缓冲队列非均衡流量性能不足的问题.

关键词: 交换结构; 负载均衡; 交叉点缓冲; 服务质量
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2012)12-2360-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.12.002

A New Type of Load-Balanced Crosspoint-Queued Switch Fabric

XU Ning, YU Shao-hua, WANG Xue-shun
(1. *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*;
2. *Wuhan Research Institute of Posts and Telecoms, State Key Laboratory for New Optical Communication Technologies and Networks, Wuhan, Hubei* 430074, *China*)

Abstract: The combined-input-crosspoint-queued switch is constrained by the delay of flow control and speedup for output-queued Switch emulation. A pure crosspoint-queued switch does not achieve good throughput performance under non-uniform traffic pattern. A type of switch architecture, load-Balanced crosspoint-queued switch, is proposed. By a load-balanced processing with predetermined patterns at extra switch fabric stage, the non-uniform traffic arriving at input port is transformed into uniform traffic and distributed evenly to every small cross-point buffer at switch fabric and corresponding to the same input port, which can lead to simpler scheduling and better QoS performance. Analysis shows the stability of such architecture and numerical results show that this switch architecture achieves approximately delay performance as output-queued switch without internal acceleration and much better throughput than pure crosspoint-queued switch under non-uniform traffic.

Key words: switch fabric; load-balanced; crosspoint-queued; quality of service

1 引言

当前各种宽带应用的不断驱动下,互连网络正朝着高速和多媒体综合传输的方向发展.作为构建高性能宽带网络的基础,路由交换设备必须能够支持巨大的带宽容量和多种业务的服务质量(QoS)保障.路由交换设备的核心是交换架构,从更广义上说交换架构是由物理上的交换架构和相应的调度算法组成,交换架构的特性直接决定了路由交换设备的性能和服务支持能力,因此一直是网络领域的研究热点之一.

早期的交换架构主要是输出队列交换或者共享存储交换结构,由于输出端口直接调度进入输出缓冲区的数据包而且不同输出端口之间的数据流相互隔离,因此具有调度算法简单而且易于提供服务质量支持的优点.带虚拟输出队列(VOQ)的 IQ 结构被证明[1]能够提供 100%的吞吐量而且不需要任何内部加速,IQ 交换结构需要全局的调度算法,随着线速率以及端口数目的增长,全局调度算法的复杂度使得线速转发决策变得越来越困难.因此寻找调度算法简单、不需要内部加速的新的交换结构以满足不断发展的网络交换带宽需要,成为了一个亟待解决的问题.

收稿日期:2011-12-09;修回日期:2012-02-27
基金项目:国家高技术研究发展计划"八六三"项目(No.2005AA121410)

2　相关研究背景

近来一种新的交换结构:混合输入-交叉点缓冲排队(CICQ)交换结构引起了广泛的注意,如图 1 所示,CICQ 交换机结构混合了输入线卡上的大容量的虚拟输出队列和交换矩阵交叉点上的缓冲,由于交叉点缓冲有效的实现了输入和输出调度的隔离,因此可以在 N 个输入调度器和 N 个输出调度器上分别采用非全局的调度算法.

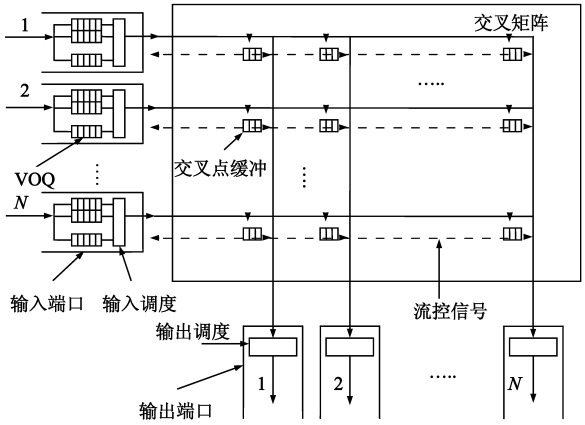

图1　CICQ交换结构

基于这种较为简单的交换结构,有很多高效的调度算法被提了出来,例如:LQF-RR[2]、SCBF[3],MCBF[4]等等,这些算法显示了比目前实用的 IQ 算法更好的吞吐量、延时等方面的性能,而且由于采用局部优化调度算法,因此也具有较低的复杂度.文献[5]证明了 CICQ 交换结构可以在 2 倍的内部加速下模拟 OQ,因此也具备与 OQ 类似的 QoS 能力.文献[6,7]提出了在 CICQ 架构下实现公平调度的算法.文献[8,9]提出了基于流的 CICQ 调度算法.

尽管具有高效和低复杂度的特性,CICQ 交换结构仍然存在一些难以克服的缺点,首先是输入线卡和交换矩阵之间通信的问题.典型的线卡和交叉点之间信息往返延时会达到 600ns[10],而在 40Gbps 的线速率下,一个 64 字节数据包占用的时隙不过 8ns 而已,此时线卡和 Crossbar 之间流控的通信开销显然已经严重影响了延时性能.此外,CICQ 仍然需要至少 2 倍交换加速来模拟 OQ 交换[11],对于最新交换网络中线速率高达 40Gbps 乃至 100Gbps 的端口速率而言,2 倍交换加速给系统的复杂性带来了很大的影响.

CICQ 比传统的 IQ 结构性能得以提高的关键在于,采用了交叉点缓冲,使得输入和输出能够分别调度不同输入/输出端口对之间的数据流,从而简化了调度、提高性能.那么能否取消掉输入线卡上的虚拟输出队列,而让数据包直接进入交叉点缓冲,在交叉点缓冲中排队调度,这样可以取消输入线卡和交换矩阵之间调度通信的开销,而且数据包进入交叉点缓冲后可以由输出调度器直接调度,进一步提高了延时性能.从技术上说这样做是可行的,因为集成电路技术的发展使得在单个芯片上集成大容量分布式缓冲成为可能,根据文献[12],片上 SRAM 单元的集成度以每 2.5 年一倍的速度增长,交叉点缓冲的容量也会随之而增长.最近的文献[13]提出了一个简单的交叉点缓冲(Crosspoint Queued,CQ)交换架构,但是仿真和分析都表明,它在非均衡流量下的吞吐量性能不好,对于某些简单的调度算法(例如轮询)来说是不稳定的(无法达到 100% 的吞吐量),在这方面的性能仍然比不上 CICQ 结构.

3　负载均衡交叉点缓冲交换结构

本文中的交换结构都是 $N \times N$ 的交换机,N 为输入/输出的端口数量.为了便于操作和理论分析,交换的时间会被划分为定长的时隙.在任意一个时隙内,只能有一个定长信元从某一个输入端口到另一个输出端口,因此也就不存在内部加速.

考虑到片上分布式缓冲的技术限制,交叉点的分布式缓冲队列容量仍然不能与线卡上的输入缓冲队列相比,纯交叉点缓冲交换结构在非均衡的流量下容易造成交叉点缓冲中的重载队列溢出,特别是在运行某些简单的调度算法(例如 Round-Robin)下时,重载的交叉点缓冲队列的丢包就会更加频繁,这会严重的影响交换结构的最大吞吐量.

显然,问题的症结在于纯交叉点缓冲结构中交叉点缓冲的容量较小,不能像输入缓冲队列一样容纳大量的突发非均衡流量,使得输出调度的性能下降.进一步的分析表明,虽然在交叉点缓冲交换结构中每个交叉点的缓冲容量有限,但是每一个输出端口都对应 N 个交叉点缓冲,因此每对输入/输出端口可以使用的缓冲区总量并不少,只是在纯交叉点缓冲结构的限制下,每一个输入只能使用输出端口对应的 N 个交叉点缓冲的其中一个,即使输出端口对应的其他的输入端口的缓冲区空闲也不能得到有效的利用,此外文献[13]的实验结果也表明,在均衡流量下,也就是每个交叉点缓冲的流量模式相同的情况下,纯交叉点缓冲能够达到较好的性能.如果能够找到一种简单而有效的方法,使从输入端口 i 到输出端口 j 的流量能够访问输出端口 j 对应的所有 N 个交叉点缓冲,而且能够使得每个交叉点缓冲的到达过程都能够近似于均衡流量,就能够克服单纯交叉点缓冲交换的缺点,更好的发挥交叉点缓冲矩阵调度过程简单且不需要输入缓冲队列和流控的优势.

基于以上分析,本文提出了一种新的交叉点缓冲交换结构——负载均衡交叉点缓冲交换(LB-CQ),其基

本的思路是,在简单交叉点缓冲交换矩阵的前面,添加一级负载均衡交换矩阵,负载均衡交换矩阵通过简单的端口轮转匹配对输入端口到达流量进行负载均衡分配,可以将到达交换矩阵各输入端口的非均衡流量通过负载均衡转化为近似均衡的流量,从而在保留 CQ 交换矩阵调度算法简单、不需要流控、延时等 QoS 性能出色的基础上,克服 CQ 交换矩阵非均衡流量下吞吐量性能不佳的缺点.

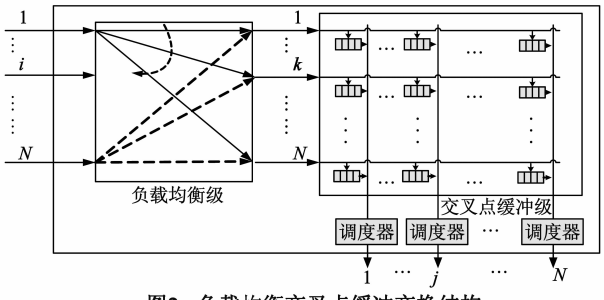

图2 负载均衡交叉点缓冲交换结构

负载均衡交叉点缓冲交换结构由两级交换矩阵组成,如图 2 所示,其中第一级是一个无缓冲的 $N \times N$ 交换矩阵,采用最简单的定时轮转模式匹配输入、输出端口,在任意时隙 t 到达第一级交换矩阵的输入端口 i 的交换信元通过预定模式的当前时隙交换节点配置,立即交换到第二级的某个中间输入端口 k,而第一级的输入、输出端口的配对在连续 N 个时隙中是不同的,例如第一级的输入端口 i 和中间输入端口 k 在第 t 个时隙的匹配模式可以设置为:

$$k = (i + t) \bmod N \tag{1}$$

也就是说,在 t 时刻到达输入端口 i 的交换信元将被立即送到第 $i + t$ 个中间输入端口,在 N 个时隙内,输入端口 i 将会轮流访问所有的中间输入端口(1..k..N),当有突发的大批信元到达同一个输入端口时,它们将被分配到不同的中间端口,然后进入第二级交换矩阵的不同交叉点缓冲,从而达到充分利用第二级交换矩阵中每个输入端口对应的 N 个交叉点缓冲容量的目的.在这样一个定时轮转交换模式下,第一级输入端口 i 每隔 N 时隙才能访问一次特定的中间输入端口 k(k = 1,2…N),且在 N 个时隙中所有的输入端口都将轮流访问一次中间输入端口 k,每个输入端口到达的流量被平均分配到了 N 个中间输入端口,且由于第一级轮转交换的对称性每个中间输入端口到达的流量等于所有输入端口到达流量的算数平均值,因而每个中间输入端口的到达流量相等,实现了将输入端口上的非均衡到达过程转化为中间输入端口上的近似均衡到达过程,后面将通过理论分析证明这一点.第二级交换矩阵是一个交叉点缓冲交换矩阵 Q,从中间输入端口 k 到达第二级交换的信元会直接送到相应的交叉点缓冲 B_{kj},输出端口 j 上的调度器在每个时隙可以从对应于 j 的一列交叉点缓冲中选出合适的信元,将其调度到输出端口上输出.

通过增加一级负载均衡交换矩阵,实现了将输入端口的突发或非均衡流量转化为中间输入端口上的近似均衡流量,解决了单纯交叉点缓冲非均衡流量性能下降的问题,交换付出的代价仅仅是增加一级简单的定时轮转交换矩阵,交换节点的个数从 N^2 增加到 $2N^2$,新增加的交换矩阵的调度算法复杂度为 $O(1)$不需要缓冲,因此不会使用比纯交叉点缓冲交换更多的分布式存储器.获得的好处是:(1)与 CICQ 交换结构相比,取消了输入缓冲队列和输入调度,消除了影响性能的流控延时,明显提高了交换结构的延时特性.(2)由于数据包直接进入交叉点缓冲,不需要进行输入调度和输入缓冲,因此可以通过交叉点的输出调度直接模拟输出排队交换机,获得最佳的 QoS 性能.(3)与纯交叉点缓冲交换相比,由于负载均衡级交换将突发和非均衡流量转换为进入纯交叉点缓冲交换矩阵的近似均衡的流量,因此在非均衡和突发流量模式下可以获得更好的性能.(4)与纯交叉点缓冲交换相比,同样的输出调度算法下需要更少的交叉点缓冲容量,减少了成本.

4 LB-CQ 的性能分析

4.1 吞吐量性能

考虑到负载均衡级既没有缓冲,输入到输出端口之间也没有任何冲突,所以影响吞吐量的主要是交叉点缓冲交换的调度,用 $A_{i,j}(n)$表示到时隙 n 为止,到达交换机的输入端口 i,目的地为输出端口 j 的数据包积累数量.用 $B_{k,j}(n)$表示到达第二级交换的中间输入端口 k,目的地为输出端口 j 的信元积累数量,用 $D_{k,j}(n)$表示从输出端口 j 离开的来自中间输入端口 k 的信元积累数量.用 $X_{k,j}(n)$表示 n 时隙交叉点缓冲的队列长度.在任意输入端口 i 到达并且目的地为任意输出端口 j 的信元满足平均速率为 $\lambda_{i,j}$的伯努利分布.若输入/输出端口对之间的流量满足 $\sum_i \lambda_{ij} \leqslant 1$ 以及 $\sum_j \lambda_{ij} \leqslant 1$,称之为可接受(Admissible)速率.

定义 1 若一个交换调度算法在其调度的缓冲区非空时,总是能够从缓冲区调度信元到输出端口,则称之为尽力而为的.

定义 2 工作在某个特定调度算法下的交换结构对于任意具有可接受速率 $\lambda_{i,j}$的到达过程 $A_{i,j}(n)$,$\lim\limits_{n\to\infty} \dfrac{D_{ij}}{n} = \lambda_{ij}, \forall\, i,j = 1,\cdots,N$ 成立,则称该交换结构速率稳定.这意味着该交换结构在此调度算法控制下能够达到 100% 的吞吐量.

根据文献[2]给出的流体模型,LB-CQ 交换排队系统的流体模型可以由以下方程描述

$$X_{kj}(t) = X_{kj}(0) + \lambda_{kj}t - D_{kj}(t) \geqslant 0 \tag{2}$$

$$D'_{kj}(t) = 1 \text{ 若 } X_{kj}(t) \geqslant 0 \text{ 且 } \pi_{kj} \in \Pi(t) \tag{3}$$

其中 t 为连续的时间,$\Pi(t)$ 为 t 时刻调度算法产生的匹配,$D'(t)$ 为 $D(t)$ 的导数.方程成立的条件是中间级端口到输出端口的速率满足强大数定理且是可接受的.

定义 3[2]　工作在某个特定调度算法下的交换结构,若对于每个流体模型解(D,Π,X)(初始条件 $X(0)=0$),有 $X(t)=0$,对于 $t \geqslant 0$ 几乎处处成立,其流体模型被称为弱稳定的.

定理 1　若交叉点缓冲交换调度采用的是尽力而为的调度算法.则 LB-CQ 交换结构在可接受且各态遍历的伯努利到达过程下是速率稳定的.

证明:首先证明到达任意中间输入端口 k 到任意输出端口 j 的流量速率满足强大数定律并且是可接受的,由于负载均衡级的均衡算法,从 i 到 j 的流量每隔 N 个时隙才能送到中间输入端口 k,因此经中间端口 k 的从 i 到 j 伯努利流量可以近似为均值 $\lambda_{ikj}=\lambda_{ij}/N$[16]的伯努利到达过程 b_{ikj},且不同的输入端口到达中间级的过程是相互独立的,因此中间端口 k 的平均速率满足

$$\lambda_{kj} = \mathrm{E}[b_{kj}] = \mathrm{E}[\sum_i b_{ikj}] = \sum_i \mathrm{E}[b_{ikj}] = \sum_i \lambda_{ij}/N \tag{4}$$

由到达过程的各态遍历性可知:

$$\lim_{n\to\infty}\frac{B_{kj}(n)}{n} = E[b_{kj}] = \lambda_{kj}$$

中间输入端口的到达过程满足强大数定理且满足

$$\sum_k \lambda_{kj} = \sum_i \lambda_{ij} \leqslant 1,\ \sum_j \lambda_{kj} = \sum_i \sum_j \lambda_{ij}/N \leqslant 1$$

因此符合文献[2]中应用流体模型方程的前提条件.

现在证明 LB-CQ 交换排队系统的流体模型满足弱稳定条件:

由于 $D'_{kj}(t)$ 与具体的调度算法有关,为了得到一个更一般性的结果,且对应同一个端口的 N 个交叉缓冲由同一个调度器调度,可以视为输出端口的虚拟输入队列,因此可以将对应每个输出端口的 N 个交叉点缓冲作为一个整体来分析.

令 $X_j(t) = \sum_k X_{kj}(t)$

显然若 $X_j(t)$ 满足弱稳定条件($X(t)=0$,对于 $t\geqslant 0$ 几乎处处成立),因为

$$\sum_k X_{kj}(t) \geqslant \max\{X_{kj}(t)\},$$

则 $X_{kj}(t)$ 也满足弱稳定条件.

LB-CQ 的流体方程可以改写为:

$$X_j(t) = X_j(0) + \sum_k \lambda_{kj}t - D_j(t) \geqslant 0 \tag{5}$$

显然有 $D'_j(t)=1$,若 $X_j(t)\geqslant 0$ 且调度算法尽力而为.

则当 $X_j(t)\geqslant 0$,有

$$X'_j(t) = \sum_k \lambda_{kj} - 1 \leqslant 0 \tag{6}$$

根据文献[11]引理 1,若 $X_j(t)\geqslant 0$ 时 $X'_j(t)\leqslant 0$,则 $X_j(t)$ 满足定义 3,即流体模型(D,Π,X)是弱稳定的,又由文献[11]定理 3,若流体模型弱稳定,对应的交换系统为速率稳定,即 LB-CQ 系统能够达到 100% 吞吐量.

定理得证.

定理 1 说明,采用 LB-CQ 交换结构是可行的而且是速率稳定的,能够满足现代交换系统的吞吐量性能要求.

4.2　模拟输出缓冲交换机

输出缓冲交换机能够提供最好的速率和服务质量保证,在此用"计数"[15]的方法来证明 LB-CQ 交换机能够在不需要内部加速的情况下模拟 Push In First Out(PIFO)输出缓冲交换机,其中 PIFO 是一类缓冲队列策略的总称,其中最典型的就是加权公平队列调度算法(WFQ).为了描述的方便,首先将每一个时隙划分为三个阶段:到达、负载均衡、输出调度.

为了能够模拟输出缓冲交换机(OQ),首先定义一个虚拟的输出缓冲交换机,它能够感知输入端口和交叉点缓冲的所有信元,并且按照输出排队的调度算法决定它们的输出顺序和离开时间.而 LB-CQ 为了模拟 OQ 的输出过程,在每个输出端口的调度器上维持一个对应的 N 个缓冲中信元的虚拟优先级队列,每当信元到达交叉点缓冲时,输出调度器就把它们放入虚拟优先级队列的对应位置.为了描述证明过程,采用来自文献[5]的定义:

定义 4　输出缓冲(Output Cushion, OC):信元的输出缓冲 $OC(p)$ 是指在虚拟 OQ 的调度下,输出队列中离开时间小于信元 p 的信元个数.

定义 5　输入线程(Input Thread, IT):输入线程 $IT(p)$ 是指在输入端口上的优先级列表中,比信元 p 优先级更高的信元个数.

定义 6　松弛度(Slackness):$L(p)$ 表示当前时刻信元 p 的输出缓冲和输入线程之差,它反映了信元需要从输入端口被传输到输出端口的紧迫度.当信元离开输入端口,$L(p)=0$.

定理 2　LB-CQ 能够在不需要内部加速的情况下能够模拟 PIFO-OQ 交换机,不论输入为何种流量模式.

证明:采用数学归纳法,在 $n=0$ 时刻,若没有信元到达,显然 LB-CQ 能够模拟 OQ,若有信元到达,此时则

信元会被直接送到交叉点缓冲,从而能够被调度到输出端口,因此 OQ 可以被模拟.假定 LB-CQ 交换机在[0, $n-1$]时隙中成功的模拟了 PIFO-OQ 交换机,若 n 时隙没有信元到达,因为所有需要调度的信元在上一个时隙都已经进入了输出端口对应的交叉点缓冲,可以由调度器进行调度,则 LB-CQ 仍然模拟了 OQ.若 n 时隙的到达阶段中,信元 p 到达输入端口,此时 $IT(p)=1$,随即信元会在 n 时隙的负载均衡阶段被负载均衡级直接传输到中间输入端口并被送入相应输出端口对应的交叉缓冲队列,此时 $IT(p)=0, L(p)=0$,说明信元已经可以由输出端口调度,在 n 时隙的调度阶段有两种情况:(1)$OC(p)=0$,此时信元立即被输出调度器调度在下一时隙输出.(2)$OC(p)>0$,此时信元会被放入输出调度的虚拟输出队列中,等待 $OC(p)=0$,即所有比 p 高优先级的信元都输出之后再输出 p.两种情况下,LB-CQ 对信元的处理方式都与输出排队缓冲交换机一致.

定理证毕.

前面提到,CICQ 交换结构需要 2 倍的交换加速才能够做到模拟 OQ,即使是最新的采用负载均衡 CICQ 的方案[17],虽然也具有从一个输入端口访问输出端口对应的 N 个交叉点缓冲的能力,但是由于没有取消输入调度和输入缓冲排队,仍然需要在一个时隙内进行两次缓冲-交换过程,才能保证信元进入输出端口,也就是仍然需要两倍内部交换加速才能达到 OQ 的性能.根据定理 2,采用 LB-CQ 交换结构可以不需要交换矩阵加速直接模拟 OQ,显然采用负载均衡的 CQ 交换结构可以获得比负载均衡 CICQ 交换结构更好的性能.

5 实验结果

为了验证 LB-CQ 交换结构的性能,采用了在 Stanford 大学开发的 SIM 交换仿真平台[18](该平台本身只适用于仿真输入排队(IQ)交换)基础上自行开发的通用交换结构仿真平台 ESIM,在这个平台上实现了 LB-CQ(采用 LQF 和 Round-Robin 调度算法)以及作为对照的 CICQ 交换机(基于 LQF-RR[2]、LB_ CICQ-FA[17]、MCBF[4] 算法),同时还实现了输出排队(OQ)交换和交叉点缓冲交换(CQ).仿真的交换结构都由 32 个输入端口和 32 个输出端口构成,每个信元的包长固定为 64 字节,占用一个时隙的时间.为了更好地模拟真实网络中流量突发特性给交换结构带来的挑战,采用中断伯努利到达过程来(IBP)模拟每个端口生成流量的概率分布,中断伯努利过程是一个 ON-OFF 过程,可以用来模拟变长数据包的突发流量,选择平均突发长度为 15 个时隙(等同于变长数据包的平均包长 15 个信元),平均端口负载 λ 的大小可以通过调整突发间隔的平均长度来设定,分成两种不同的流量场景,均衡的($\lambda_{ij}=\lambda/N$)以及非均衡的($\lambda_{ii}=0.7\lambda$, $\lambda_{i|i+1|}=0.3\lambda$,其他输入输出端口对 = 0).为了得到稳定的结果,每次仿真每种流量设定都至少仿真了 3×10^6 个时隙,仿真重复进行了 5 次.

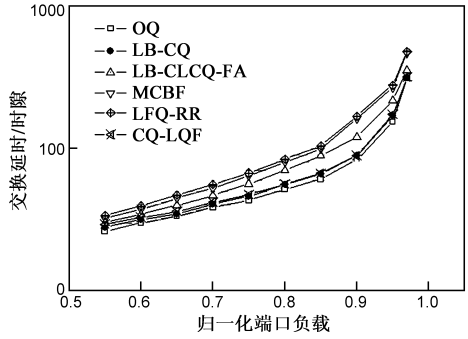

图3 均衡突发流量下的平均交换延时

图 3 说明了在均衡突发流量下不同交换结构以及调度算法的交换延时性能,如图所示,很明显 LB-CQ 的延时性能优于所有的 CICQ 交换结构,即使同样采用负载均衡机制的 CICQ 也不例外,而且在不考虑流控信号传输造成的延时情况下,如果考虑流控延时,则 LB-CQ 的性能优势将更加明显.在均衡流量下 LB-CQ 的延时性能与 CQ 近似且与 OQ 的差别不大.

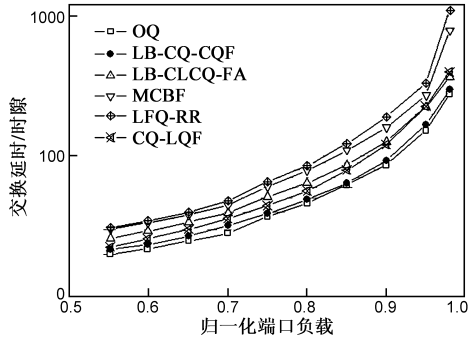

图4 非均衡突发流量下的平均交换延时

图 4 说明了在非均衡流量下不同交换结构以及调度算法的交换延时性能,从仿真结果可以清楚的显示在非均衡流量下,LB-CQ 仍然具有超过包括采用负载均衡在内的 CICQ 交换结构的延时性能,而此时 CQ 的延时性能随着吞吐量的增加而开始明显下降,LB-CQ 与 OQ 在非均衡流量下的性能差别仍然不大,这也证实了对于性能的分析.然而 OQ 为了达到这样的延时性能付出了 N 倍加速的代价.

为了进一步说明 LB-CQ 比类似于文献[13]中的 CQ 更好的适应非均衡流量,采用非均衡 log diagonal($\lambda_{ij}=2\lambda_{i|j+1|}$且 $\sum_i \lambda_{ij}=1$)流量模式仿真了这两种交换结构在相同的交叉点缓冲容量下采用轮询调度(Round-Robin)所能够达到的最大吞吐量.仿真结果如图 5 所示,在非均衡 log diagonal 流量下,LB-CQ 即便采用容量不大的交叉点缓冲,仍然能够达到 100% 的吞吐量.而

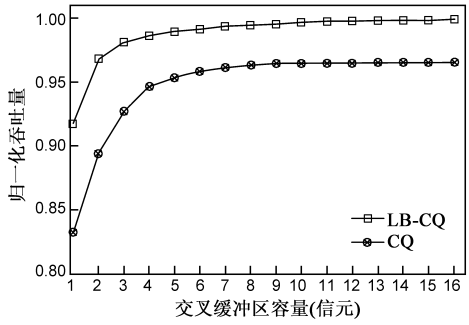

图5　LB-CQ和CQ在非均衡流量下最大吞吐量

CQ 在非均衡流量下的吞吐量性能则存在明显不足,即使不断加大交叉点缓冲的容量,也无法使 CQ 在轮询调度算法下达到 100%的吞吐量.

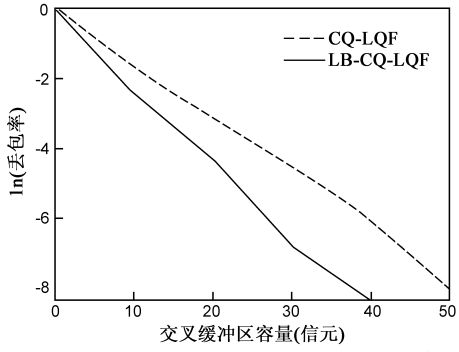

图6　LB-CQ和CQ在非均衡流量下丢包率

单纯的 CQ 交换结构即使采用全局优化调度算法,例如 LQF,仍然需要较大的交叉缓冲来避免丢包的产生,LB-CQ 对交叉点缓冲的要求要小得多.图 6 显示了采用 LQF 调度算法在非均衡突发流量下 CQ 和 LB-CQ 的丢包率与交叉点缓冲区规模的关系.实验表明,采用相同规模的交叉点缓冲,LB-CQ 能达到更低的丢包率.说明 LB-CQ 能够用较小规模的缓冲区达到同样的丢包性能,从而降低了实现的难度.

6　结论

CICQ 交换结构最大的优点是通过交叉点缓冲,实现了输入调度和输出调度的解耦合,因而简化了调度实现的复杂度并且获得了较好的性能.将这一优点发挥到极致的结果就是纯粹的交叉点缓冲交换结构(CQ),CQ 交换结构完全取消了输入调度步骤,进一步简化了调度算法,同时使得延时等方面的性能得到进一步提高.但是,单纯的 CQ 交换结构在非均衡流量模式下吞吐量不能令人满意.本文提出了新的交换结构——负载均衡交叉点缓冲交换结构(LB-CQ),通过负载均衡将输入端口的非均衡流量变换为均衡流量,从而解决了 CQ 在非均衡流量下吞吐量性能不足的缺陷并且保留了 CQ 交换结构的低延时特性.分析和仿真都表明这种新型的交换结构在性能接近于 OQ 交换结构的同时,其非均衡流量下的吞吐量、丢包率等性能比 CQ 有明显的提高.下一步的主要工作将集中在如何更有效的利用分布式交叉缓冲从而在保持 CQ 的优点的同时进一步降低实现的难度.

参考文献

[1] McKeown N, Mekkittikul A, Anantharam V. Achieving 100% throughput in an input-queued switch[J]. IEEE Transactions on Communications, 1999, 47(8): 1260 – 1267.

[2] Javidi T, Magill R, Hrabik T. A high-throughput scheduling algorithm for a buffered crossbar switch fabric[A]. IEEE ICC'01 [C]. St Petersburg, Russia : IEEE, 2001. 1581 – 1587.

[3] Lotfi Mhamdi, Mounir Hamdi. CBF: A high-performance scheduling algorithm for buffered crossbar switches[A]. 4th High Performance Switching and Routing(HPSR '03)[C]. Torino, Italy : IEEE, 2003. 67 – 72.

[4] Zhang X and Bhuyan L N. An efficient algorithm for combined input-crosspoint-queued (CICQ) switches[A]. IEEE Globecom '04[C]. Dallas, USA: IEEE, 2004. 1168 – 1173.

[5] Chuang S T, Iyer S, McKeown N. Practical algorithms for performance guarantees in buffered crossbars[A]. IEEE INFOCOM'05[C]. Miami, USA : IEEE, 2005. 981 – 991.

[6] Zhang X, Mohanty S R, Bhuyan L N. Adaptive max-min fair scheduling in buffered crossbar switches without speedup[A]. IEEE INFOCOM'07[C]. Anchorage, USA: IEEE, 2007. 454 – 462.

[7] Hosaagrahara M, Sethu H. Max-min fair scheduling in input-queued switches[J]. IEEE Transactions on Pararrel and Distribute Systems, 2008, 19(4): 462 – 475.

[8] Szymanski T H. A Low-jitter guaranteed-rate scheduling algorithm for crosspoint-buffered switches[A]. Communications, Computers and Signal Processing 2009 [C]. Victoria, BC: IEEE, 2009. 882 – 890.

[9] Divakaran D M, Anhalt F, Altman E. Size-based flow scheduling in a CICQ switch[A]. 11th High Performance Switching and Routing(HPSR'10)[C]. Dallas, USA: IEEE, 2010. 57 – 62.

[10] Minkenberg A F, Iliadis C I, Engbersen A P. Design issues in next-generation merchant switch fabrics [J]. IEEE/ACM Transactions on Networking, 2007, 15(6): 1603 – 1615.

[11] Dai J G, Prabhakar B. The throughput of data switches with and without speedup[A]. IEEE INFOCOM 2000[C]. Tel-A-viv, Israel: IEEE, 2000. 556 – 564.

[12] ITRS Committee. International Technology Roadmap for Semiconductors Executive Summary[R]. New York: ITRS, 2007. 13 – 17.

[13] Kanizo Y, Hay D, Keslassy I. The crosspoint-queued switch [A]. IEEE INFOCOM'09. [C]. Janeiro, Brail: IEEE, 2009. 729 – 737.

[14] Shen Y, Panwar S S, Chao H J. Providing 100% throughput in a buffered crossbar switch[A]. 8th High Performance Switching and Routing (HPSR' 07)[C]. New York, USA: IEEE, 2007. 1 - 8.

[15] Magill R B, Rohrs C E. Output-queued switch emulation by fabrics with limited memory[J]. IEEE Journal on Selected Areas in Communications, 2003, 21(4): 606 - 615.

[16] Shen Y, Panwar S S, Chao H J. Design and performance analysis of a practical load-balanced switch[J]. IEEE Transactions on Communications, 2009, 57(8): 2420 - 2429.

[17] Rojas-Cessa R, Dong Z. Load-balanced combined input-crosspoint buffered packet switches [J]. IEEE Transactions on Communications, 2011, 59(5): 1421 - 1433.

[18] McKeown N. SIM [EB/OL]. http://klamath.stanford.edu/tools/SIM/, 2007-05-16/2009-07-11.

[19] 李挥, 何伟, 伊鹏, 王秉睿, 雷凯, 安辉耀, 汪斌强. 排序集线器多级互连交换结构的多路径自路由模型[J]. 电子学报, 2008, 36(1): 1 - 8.
Li Hui, He Wei, Yi Peng, Wang Bing rui, Lei Kai, An Hui yao, Wang Bin qiang. Modeling multi-path self-routing switching structure from multistage interconnection of sorting concentrators [J]. Acta Electronica Sinica, 2008, 36(1): 1 - 8. (in Chinese)

作者简介

徐 宁 男, 1976 年 11 月生于武汉, 博士研究生, 研究方向: 数据交换结束、IP 互联网络.
E-mail: nxu@smail.hust.edu.cn

余少华 男, 1962 年 5 月生于武汉, 博士, 教授级高工, 博士生导师, 主要研究方向为 IP 互联网络和城域网.
E-mail: shyu@fhn.com.cn